AD 684 428

[illegible]

[illegible] STATISTICAL RESEARCH PROGRAM

[illegible] VARIABLES

by

[illegible]

Technical Report No. 26
Office of Naval Research THEMIS Contract

Distribution of this document is unlimited.

Department of Statistics
Southern Methodist University
Dallas, Texas 75222

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

CORRELATION BETWEEN TWO VECTOR VARIABLES

by

A. M. Kshirsagar

Technical Report No. 26
Department of Statistics THEMIS Contract

March 4, 1969

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

CORRELATION BETWEEN TWO VECTOR VARIABLES

by

A. M. Kshirsagar*
Southern Methodist University
Dallas, Texas 75222

SUMMARY

H. Ruben (1966) has suggested a simple approximate normalization for the correlation coefficient in normal samples, by representing it as the ratio of a linear combination of a standard normal variable and a chi variable to an independent chi variable and then using Fisher's approximation to a chi variable. This result is extended in this paper to a matrix, which in a sense is the correlation coefficient between two vector variables $\underset{\sim}{x}$ and $\underset{\sim}{y}$. The result is then used to obtain large sample null and non-null (but in the linear case) distributions of the Hotelling-Lawley criterion and the Pillai criterion in multivariate analysis. Williams (1955) and Bartlett (1951) have derived some exact tests for the goodness of fit of a single hypothetical function to bring out adequately the entire relationship between two vectors $\underset{\sim}{x}$ and $\underset{\sim}{y}$, by factorizing Wilks' Λ suitably. These factors are known as "direction" and "collinearity" factors, as they refer to the direction and collinearity aspects of the null hypothesis. In this paper, the other two criteria viz. the Hotelling-Lawley and Pillai criteria are partitioned into direction and collinearity parts and large sample tests corresponding to them are derived for testing the goodness of fit of an assigned function.

*This research was sponsored by the Office of Naval Research, Contract No. N00014-68-A-0515, Project No. NR042-260. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. INTRODUCTION

If r is the sample correlation coefficient between x and y, r^2 is the ratio of the regression sum of squares to the total sum of squares and $r^2/(1 - r^2)$ is the ratio of the regression sum of squares to the residual sum of squares, in the regression of x on y. When however, we consider the regression of a $p \times 1$ vector $\underline{x}$ on a $q \times 1$ vector $\underline{y}$, $(p \leq q)$ we shall obtain two $p \times p$ symmetric matrices corresponding to regression of $\underline{x}$ on $\underline{y}$ and the residual. Let these be denoted by B and A respectively so that $A + B$ is the "total" matrix. Matrix generalizations of r^2 and $r^2/(1 - r^2)$ can be obtained from B, A and $A + B$ by expressing $A + B$ as CC' and A as FF' where C and F are lower triangular matrices. Then $C^{-1}BC'^{-1}$ can be looked upon as a generalization of r^2 and $F^{-1}BF'^{-1}$ of $r^2/(1 - r^2)$. Ruben (1966) expressed $\tilde{r} = r/\sqrt{1 - r^2}$ as

$$(\xi + \tilde{\rho}\chi_{n-1})/\chi_{n-2}$$

where ξ is a $N(0, 1)$ variable, χ_a denotes a chi-variate with 'a' degrees of freedom (d.f.) and ξ, χ_{n-1}, χ_{n-2} are independent, $\tilde{\rho}$ is the population parameter. A similar representation is derived in this paper for the matrix generalization of $\tilde{r}$ and is used to obtain an approximate large sample normalization of this matrix.

Several multivariate problems can be put into the framework of relationship between two vectors $\underline{x}$ and $\underline{y}$. The following three criteria are generally used in multivariate analysis to test lack of association between x and y :

(1) Wilks' Λ ; $\Lambda = |A|/|A + B|$

(2) Pillai's criterion $\quad \mathrm{tr}\, B(A + B)^{-1}$

(3) Hotelling-Lawley criterion $\quad \mathrm{tr}\, BA^{-1}$

Large sample null and non-null (linear case) distributions of the last two criteria are derived, using the approximate normalization of the generalization of $\tilde{r}$ and further a suitable partitioning of the two criteria, analogous to the factorization of Wilk's Λ by Bartlett (1951), for testing the goodness of fit of a single hypothetical function $\alpha_1 x_1 + \cdots + \alpha_p x_p$, is derived.

2. MATRIX GENERALIZATION OF $\tilde{r}$

Let the variance-covariance matrix of the two vectors

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \qquad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_q \end{bmatrix}$$

be

$$\Sigma = \begin{array}{c c} \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right] & \begin{array}{c} p \\ q \end{array} \\ \begin{array}{cc} p & q \end{array} & \end{array} \tag{2.1}$$

and let the matrix of corrected sum of squares (s.s.) and sum of products (s.p.) of observations in a sample on these variables be

$$S = \begin{array}{c c} \left[\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right] & \begin{array}{c} p \\ q \end{array} \\ \begin{array}{cc} p & q \end{array} & \end{array} \tag{2.2}$$

This is based on n d.f. Then we have the following matrices:

B_0 = matrix of regression coefficients $S_{12}S_{22}^{-1}$, of x on y (2.3)

B = matrix of s.s. & s.p. due to regression $S_{12}S_{22}^{-1}S_{21}$ (2.4)

A = "residual" s.s. & s.p. matrix $S_{11} - S_{12}S_{22}^{-1}S_{21} = S_{11\cdot 2}$ (2.5)

A + B = "total" matrix S_{11} (2.6)

The corresponding matrices for the population are:

$\beta = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and Σ_{11} respectively.

If $\underline{x}$ and $\underline{y}$ have a normal distribution, S will have a Wishart distribution and from that, by suitable matrix transformations, it can be shown that B_0 , S_{22} and A are independently distributed as below:

$$\text{(1) Const. } e^{-1/2 \text{ tr } \Sigma_{11\cdot 2}^{-1}(B_0 - \beta)S_{22}(B_0 - \beta)'} dB_0 \qquad (2.7)$$

$$\text{(2) Const. } |S_{22}|^{(n-q-1)/2}\exp\{-1/2 \text{ tr } \Sigma_{22}^{-1}S_{22}\}\, dS_{22} \qquad (2.8)$$

and $$\text{(3) Const. } |A|^{(n-q-p-1)/2}\exp\{-1/2 \text{ tr } \Sigma_{11\cdot 2}^{-1}A\}\, dA \qquad (2.9)$$

Thus B_0 has a normal distribution, while q × q matrix S_{22} has a Wishart distribution with n - q d.f. We shall denote the last two distributions (2.8) and (2.9) by $W_q(S_{22}|\Sigma_{22}|n)$ and $W_p(A|\Sigma_{11\cdot 2}|n-q)$. Let Φ , η , M , F , C , K be <u>lower triangular matrices</u> such that $\Sigma_{22} = \Phi\Phi'$, $\Sigma_{11\cdot 2} = \eta\eta'$, $S_{22} = MM'$, $A = FF'$, $B = KK'$ and $A + B = CC'$. Define further

$$U = \eta^{-1}(B_0 - \beta)M \qquad (2.10)$$

$$V = \Phi^{-1}M \qquad (2.11)$$

$$W = \eta^{-1}F \qquad (2.12)$$

$$\tilde{P} = \eta^{-1}\Sigma_{12}\Phi'^{-1} \qquad (2.13)$$

$$\tilde{R} = F^{-1}S_{12}M'^{-1} \qquad (2.14)$$

$$R - C^{-1}K \qquad (2.15)$$

$$L = RR' = C^{-1}BC'^{-1} \qquad (2.16)$$

It can be easily seen that $L = RR'$ is the matrix generalization of r^2 and $\tilde{R}\tilde{R}'$ is the matrix generalization of $r^2/(1 - r^2)$. Observe that

$$U + \tilde{P}V = \eta^{-1}S_{12}M'^{-1} \tag{2.17}$$

where $\tilde{P}$ is the population matrix corresponding to $\tilde{R}$. Hence

$$\tilde{R} = W^{-1}(U + \tilde{P}V) \tag{2.18}$$

and

$$\tilde{R}\tilde{R}' = F^{-1}BF'^{-1} \tag{2.19}$$

If we transform to U , V and W from B_0 , S_{22} and A respectively in (2.7), (2.8) and (2.9), it can be easily seen that

(1) u_{ij} $(i = 1 , \cdots , p; j = 1 , \cdots , q)$, the pq variables in U are independent N(0 , 1) variables.

(2) v_{jj} $(j = 1 , \cdots , q)$ are independent χ_{n-j+1} variates and the off-diagonal elements v_{kj} $(k > j , k = 1 , \cdots , q$ $j = 1 , \cdots , q)$ of the lower triangular matrix V are independent N(0 , 1) variables, independent of v_{jj} also.

(3) W_{ii} $(i = 1 , \cdots , p)$, the diagonal elements of the lower triangular matrix W are independent $\chi_{n-q-i+1}$ variates, while W_{ik} $(i > k , i , k = 1 , \cdots , p)$ are independent N(0 , 1) variables, independent of W_{ik} also.

Since a Wishart distribution is the multivariate matrix generalization of a χ^2 distribution, V or W , which are in a certain sense matrix square roots of $\Sigma_{22}^{-1}S_{22}$ and $\Sigma_{11\cdot 2}^{-1}A$ can be looked upon as matrix generalizations of a chi-variate. This is further supported by the fact that the diagonal elements of V and W are chi variables. Consequently (2.18) is the multivariate analogue of Rubin's representation

$$r = \chi_{n-2}^{-1}(\xi + \delta\chi_{n-1}) \; , \tag{2.20}$$

stated earlier.

Ruben uses Fisher's approximation of a chi-variate viz. χ_a is approximately normal with mean $(a - 1/2)^{1/2}$ and variance 1/2 and proves that

$$\frac{\left(\frac{2n-5}{2}\right)^{1/2}\tilde{r} - \left(\frac{2n-3}{2}\right)^{1/2}\tilde{\rho}}{(1 + \tilde{r}^2/2 + \tilde{\rho}^2/2)^{1/2}} \tag{2.21}$$

is approximately $N(0\ ,\ 1)$. This is a fairly good approximation for all practical purposes. We now proceed to consider a similar result for our $\tilde{R}$. Ruben derived (2.21) by equating (2.20) to $\tilde{r}_0$ and then showing that the approximate normal variate

$$\xi + \tilde{\rho}\chi_{n-1} - \chi_{n-2}\tilde{r}_0$$

has mean

$$\left(\frac{2n-3}{2}\right)^{1/2}\tilde{\rho} - \left(\frac{2n-5}{2}\right)^{1/2}\tilde{r}_0$$

and variance

$$1 + 1/2(\tilde{\rho}^2 + \tilde{r}_0^2)$$

He then replaces $\tilde{r}_0^2$ by $\tilde{r}^2$ to get (2.21) . We employ a similar procedure mechanically with the hope of obtaining a suitable approximation to the distribution of $\tilde{R}$. Consider the matrix

$$\xi = U + \tilde{P}V - W\tilde{R}_0\ , \tag{2.22}$$

where $\xi = [\xi_{ij}]$, $\tilde{R}_0 = [\tilde{r}^0_{ij}]$, $(i = 1\ , \cdots\ , p;\ j = 1\ , \cdots\ , q)$.

Using Fisher's approximation of a χ variate by a normal variate, for v_{ij}

and w_{ii} , we can see by a little algebra that the ξ_{ij} are normally distributed and

$$E(\xi_{ij}) = \left(\frac{2n-2j+1}{2}\right)^{1/2} \tilde{\rho}_{ij} - \left(\frac{2n-2q-2i+1}{2}\right)^{1/2} \tilde{r}^{0}_{ij} \tag{2.23}$$

$$V(\xi_{ij}) = 1 + \sum_{k=j}^{q} \tilde{\rho}^{2}_{ik} + \sum_{k=1}^{i} \tilde{r}^{02}_{kj} - \tilde{\rho}^{2}_{ij}/2 - \tilde{r}^{02}_{ij}/2 \tag{2.24}$$

$$\mathrm{Cov}(\xi_{ij} , \xi_{i'j'}) = \begin{cases} 0 \quad i \neq i' , j \neq j' \\ \sum_{k=1}^{i} \tilde{r}^{0}_{kj}\tilde{r}^{0}_{kj'} - (1/2)\tilde{r}^{0}_{ij}\tilde{r}^{0}_{ij'} , \\ \sum_{k=j}^{q} \tilde{\rho}_{ik}\tilde{\rho}_{i'k} - (1/2)\tilde{\rho}_{ij}\tilde{\rho}_{i'j} \end{cases} \tag{2.25}$$

Following Ruben's argument for $\tilde{r}$, we expect

$$\frac{\left(\frac{2n-2q-2i+1}{2}\right)^{1/2} \tilde{r}_{ij} - \left(\frac{2n-2j+1}{2}\right)^{1/2} \tilde{\rho}_{ij}}{\left\{1 + \sum_{k=j}^{q} \tilde{\rho}^{2}_{ik} + \sum_{k=1}^{i} \tilde{r}^{2}_{kj} - \tilde{\rho}^{2}_{ij}/2 - \tilde{r}^{2}_{ij}/2\right\}^{1/2}} \tag{2.26}$$

to be approximate N(0 , 1) variable. However, on account of (2.25), $\tilde{r}_{ij}$ are not independently distributed. For large n , the numerator of (2.26) can very well be taken as

$$\sqrt{n}\ (\tilde{r}_{ij} - \tilde{\rho}_{ij}) \tag{2.27}$$

If we consider the null case viz. $\tilde{P} = 0$, we find that $\tilde{r}_{ij}$ and $\tilde{r}_{i'j}$ ($i \neq i'$) are uncorrelated and so $\sqrt{n}\ \tilde{R}$ can be approximately regarded as a random sample from a multivariate normal distribution, with zero means and a certain covariance matrix. In the bivariate case, when $\rho = 0$, we have two normal approximations available to us for large n viz. $\sqrt{n}\,\tilde{r}$ is N(0 , 1) and the other one is

$$\sqrt{n}\,\tilde{r}/(1 + \tilde{r}^2)^{1/2} = \sqrt{n}\,r \text{ is } N(0\ ,\ 1) \tag{2.28}$$

The corresponding multivariate generalizations will be

(1) $\sqrt{n}\,\tilde{R}$ is a matrix of independent N(0 , 1) variables (2.29)

in large samples, and

(2) $\sqrt{n}\,D^{-1}\tilde{R}$ is a matrix of independent N(0 , 1) variables. (2.30)

Here $D = F^{-1}C$ is a lower triangular matrix and so

$$DD' = F^{-1}CC'F'^{-1}$$
$$= F^{-1}(A + B)F'^{-1}$$

but $I = F^{-1}AF'^{-1}$ and $\tilde{R}\tilde{R}' = F^{-1}BF'^{-1}$

and therefore,

$DD' = I + \tilde{R}\tilde{R}'$, a matrix generalization of $1 + \tilde{r}^2$ of (2.28) . (2.31)

We shall investigate (b) first. If (b) is true, we shall expect the $p \times p$ matrix

$$\Gamma = nD^{-1}\tilde{R}\tilde{R}'D'^{-1} \tag{2.32}$$

to have the distribution

$$W_p(\Gamma|I|q)d\Gamma \tag{2.33}$$

for large n. Now

$$\frac{1}{n}\Gamma = C^{-1}F\tilde{R}\tilde{R}'F'C'^{-1}$$

$$= C^{-1}BC'^{-1}$$

$$= L \quad \text{by (2.16)} \tag{2.34}$$

When $\tilde{P} = 0$, B has the $W_r(B|\Sigma_{11}|q)$ distribution and A has an independent $W_p(A|\Sigma_{11}|n-q)$ distribution. We transform from A and B to C and Γ by (2.34) and CC' = A + B , integrate out C and find that the distribution Γ is (see Kshirsagar, 1961 a)

$$\text{Const.}|\Gamma|^{(q-p-1)/2}|I - \frac{1}{n}\Gamma|^{(n-q-p-1)/2}d\Gamma \tag{2.35}$$

But as $n \rightarrow \infty$,

$$|I - \frac{1}{n}\Gamma|^{(n-q-p-1)/2} \rightarrow e^{-1/2 \text{ tr } \Gamma}$$

so that, in large samples, Γ has the Wishart distribution

$$W_p(\Gamma|I|q)d\Gamma \ , \tag{2.36}$$

as we expected in (2.33), if (b) is true. This, of course, is not a proof of (b) but it supports our conjecture about the large sample behaviour of $\sqrt{n}\ D^{-1}\tilde{R}$.

As regards (a), we observe that

$$\tilde{R}\tilde{R}' = F^{-1}BF'^{-1} \quad \text{and} \quad A = FF' \tag{2.37}$$

Transforming from A and B to F and $\tilde{R}$, we find the distribution of $\Delta = n\tilde{R}\tilde{R}'$ to be

$$\text{Const.}|\Delta|^{(q-p-1)/2}|I + \frac{1}{n}\Delta|^{-(n-q+p+1)/2}d\Delta \tag{2.38}$$

This, as $n \to \infty$, tends to

$$\text{Const.}|\Delta|^{(q-p-1)/2}\exp(-1/2 \text{ tr } \Delta)d\Delta$$

or (2.39)

$$W_p(\Delta|I|q)d\Delta ,$$

as it should if (a) is true.

So, for testing the null hypothesis $\tilde{P} = 0$ or which is the same as $\Sigma_{12} = 0$, we have two criteria

$$\text{tr } \Delta = n \text{ tr } A^{-1}B \quad \text{and} \quad \text{tr } \Gamma = n \text{ tr } (A + B)^{-1}B \tag{2.40}$$

Both of them have a χ^2 distribution with pq d.f., for large n. Both these criteria are well known in literature. tr $A^{-1}B$ is Hotelling(1951)-Lawley(1938) criterion and tr $(A + B)^{-1}B$ is Pillai's criterion (1955).

3. NON-NULL DISTRIBUTIONS OF Γ AND Δ

In many practical situations $\underline{y}$ is a vector of dummy variables representing differences among q + 1 groups or populations and one is interested in constructing discriminant functions for these groups. In this case, it is known that the number of discriminant functions is equal to the number of non-zero true canonical correlations between x and y . In particular, if ρ_1 is the only non-zero true canonical correlation and $\rho_2 , \rho_3 , \cdots , \rho_p$ are all null, the group means are collinear and only one discriminant funtion is adequate. This is the canonical variate corresponding to ρ_1 . Suppose

$$\underline{\alpha}'\underline{x} = \alpha_1 x_1 + \cdots + \alpha_p x_p \tag{3.1}$$

is an assigned or hypothetical function and we want to test its goodness of fit for discriminating among q + 1 groups. The hypothesis of goodness of fit of $\underline{\alpha}'\underline{x}$ comprises of two parts:

(I) Direction Aspect: Whether $\underline{\alpha}'\underline{x}$ agrees with the true discriminant function viz. the canonical variate corresponding to ρ_1 and

(II) Collinearity Aspect: Whether one discriminant function can be adequate at all or in other words, whether ρ_1 is the only non-zero canonical correlation.

Bartlett (1951) and Williams (1955) derived exact tests based on factorization of Wilks' Λ criterion, $|A|/|A + B|$ for this purpose. Our aim here is to derive similar tests for (I) and (II) based on the other two criteria -- Hotelling(1951)-Lawley(1938) and Pillai (1955). For this purpose, we shall derive the non-null distributions of Γ and of Δ , in the linear case, i.e., the case where $\rho_1 \neq 0$, $\rho_2 = \cdots = \rho_p = 0$. This is called linear case because the means of the q + 1 groups are collinear or lie in a space of 1 dimension.

Let $\underline{x}^*$, $\underline{y}^*$ be the vectors of the true (population) canonical variables and let the relationship between x* and x be

$$\underline{x}^* = \underline{\delta}\ \underline{x} \qquad (3.2)$$

where δ is a p × p non-singular matrix. $\underline{x}^*$ and $\underline{y}^*$ have, therefore, I_p and I_q as their variance-covariance matrices respectively and except for ρ_1 , the correlation between x_1^* and y_1^* , all other correlations are zero.

Define

$$A^* = \delta A \delta' \ , \ B^* = \delta B \delta' \ , \ C^*C^{*\prime} = A^* + B^* \ ,$$

where C^* is a lower triangular matrix. (3.3)

Then, the distribution of

$$L^* = [\ell^*_{ij}] = C^{*-1}B^*C^{*\prime -1} \ , \tag{3.4}$$

when $\underline{y}^*$ is fixed, is shown to be (Kshirsagar, 1961a)

$$\text{Const.}\Phi(\ell^*_{11} \ , \rho_1)|L^*|^{(q-p-1)/2}|I - L^*|^{(n-q-p-1)/2}dL^* \tag{3.5}$$

where

$$\Phi(\ell^*_{11} \ , \rho) = e^{-\lambda^2/2}{}_1F_1\left(\frac{n}{2} , \frac{q}{2} , \frac{\lambda^2}{2} \ell^*_{11}\right) \ , \tag{3.6}$$

and

$$\lambda^2 = \rho_1^2 \sum_{r=1}^{n} y^{*2}_{1r}/(1 - \rho_1^2) \tag{3.7}$$

As in section 2, for large n

$$|I - L^*|^{(n-q-p-1)/2}$$

can be replaced by

$$\exp\left\{- \frac{1}{2} \text{tr } \Gamma^*\right\} \tag{3.8}$$

where

$$\Gamma^* = nL^* \tag{3.9}$$

and so, Γ^* will have a non-central Wishart distribution of the linear case (Anderson, 1946) for large n . Make a further transformation

$$\Gamma^* = nL^* = S^*S^{*\prime} \tag{3.10}$$

where $S^* = [S^*_{ij}]$ is a lower triangular matrix. Then it can be readily seen that, for large n , S^{*2}_{11} is a non-central χ^2 (non-centrality parameter is λ^2) , S^{*2}_{ii} is a χ^2 with q + 1 - 2 d.f. ($i = 2 , \cdots , p$) , S^*_{ij} ($i > j ; i , j = 2 , \cdots , p$) is N(0 , 1) and all these variables are independent. The over-all criterion for testing the independence of $\underline{x}$ and $\underline{y}$ (which in this case means, all the q + 1 groups have identical means) is, as seen in section 2, tr Γ , which is the same as tr Γ^* on account of (3.3) and

$$\text{tr } \Gamma^* = (S^{*2}_{11}) + (S^{*2}_{21} + \cdots + S^{*2}_{p1}) + (\sum_{\substack{i,j=2 \\ i>j}}^{p}\sum S^{*2}_{ij}) \tag{3.11}$$

$$= \gamma_1 + \gamma_2 + \gamma_3 \text{ say}$$

Then γ_1 contains the entire non-centrality; γ_2 is a χ^2 with p - 1 d.f. and γ_3 is a χ^2 with (p - 1)(q - 1) d.f.

Let

$$S = \begin{array}{cc} \left[\begin{array}{c|c} S^*_{11} & 0 \\ \hline \underline{s}^* & S^*_2 \end{array}\right] & \begin{array}{l} 1 \\ p-1 \end{array} \\ \begin{array}{cc} 1 & p-1 \end{array} & \end{array} \tag{3.12}$$

$$B^* = K^*K^{*\prime} , \quad K^* \text{ is lower triangular} \tag{3.13}$$

$$K^* = \left[\begin{array}{c|c} K^*_{11} & 0 \\ \hline \underline{k}^* & K^*_2 \end{array}\right] \tag{3.14}$$

$$C^* = \left[\begin{array}{c|c} C^*_{11} & 0 \\ \hline \underline{c}^* & C^*_2 \end{array}\right] \tag{3.15}$$

then

$$S^* = nC^{*-1}K^* , \tag{3.16}$$

and after a little algebra, we find that

$$\frac{1}{n}\gamma_1 = \frac{\underline{\delta}'_{(1)} B \underline{\delta}_{(1)}}{\underline{\delta}'_{(1)} (A + B) \underline{\delta}_{(1)}} \tag{3.17}$$

$$\frac{1}{n}\gamma_2 = \frac{\underline{\delta}'_{(1)} B (A + B)^{-1} B \underline{\delta}_{(1)}}{\underline{\delta}'_{(1)} B \underline{\delta}_{(1)}} - \frac{\underline{\delta}'_{(1)} B \underline{\delta}_{(1)}}{\underline{\delta}'_{(1)} (B + A) \underline{\delta}_{(1)}} \tag{3.18}$$

$$\frac{1}{n}\gamma_3 = \text{tr } B(A + B)^{-1} - \frac{\underline{\delta}'_{(1)} B (A + B)^{-1} B \underline{\delta}_{(1)}}{\underline{\delta}'_{(1)} B \underline{\delta}_{(1)}} \tag{3.19}$$

where $\underline{\delta}'_{(1)}$ is the first row of the matrix δ , defined by (3.2). If we are testing the goodness of fit of a hypothetical function $\underline{\alpha}'\underline{x}$, we are testing the hypothesis:

$$H: \ \rho_1 \neq 0 \ , \ \rho_2 = \cdots = \rho_p = 0 \ \text{ and } \underline{\alpha}'\underline{x} \text{ is the first true canonical variate, i.e., } x_1^* = \underline{\alpha}'\underline{x} \tag{3.20}$$

But $x_1^* = \underline{\delta}'_{(1)}\underline{x}$ by (3.2) and so, if H is true, $\underline{\alpha}$ and $\underline{\delta}_{(1)}$ are identical and hence we can use γ_2 given by (3.18) and γ_3 given by (3.19), with $\underline{\delta}_{(1)}$ replaced by $\underline{\alpha}$ for testing the "direction" aspect and the "collinearity" aspect of H. The over-all test of H is given by $\gamma_2 + \gamma_3$ and γ_2 , γ_3 are the direction and collinearity parts of tr $B(A + B)^{-1}$. The latter can be justified by an argument similar to the one employed by the author elsewhere (1961b), for testing the goodness of fit of a hypothetical principal component.

In exactly a similar manner, we can show that, for the other criterion tr BA^{-1} , the partitioning is

$$n \text{ tr } BA^{-1} = \xi_1 + \xi_2 + \xi_3 \ , \tag{3.21}$$

where

$$\frac{1}{n}\,\xi_1 = \frac{\underline{\alpha}'B\underline{\alpha}}{\underline{\alpha}'A\underline{\alpha}} \tag{3.22}$$

$$\frac{1}{n}\,\xi_2 = \frac{\underline{\alpha}'BA^{-1}B\underline{\alpha}}{\underline{\alpha}'B\underline{\alpha}} - \frac{\underline{\alpha}'B\underline{\alpha}}{\underline{\alpha}'A\underline{\alpha}} \tag{3.23}$$

$$\frac{1}{n}\,\xi_3 = \mathrm{tr}\ BA^{-1} - \frac{\underline{\alpha}'BA^{-1}B\underline{\alpha}}{\underline{\alpha}'B\underline{\alpha}} \tag{3.24}$$

ξ_2 is a χ_2 with (p-1) d.f. and ξ_3 is a χ^2 with (p-1)(q-1) d.f. in large samples and these are respectively the "direction" and "collinearity" parts and can be used to test these aspects of the null hypothesis H.

REFERENCES

Anderson, T. W. (1946). "The non-central Wishart distribution and certain problems of multivariate statistics," *Annals of Mathematical Statistics*, 17, 409-431.

Bartlett, M. S. (1951). "The goodness of fit of a single hypothetical discriminant function in the case of several groups," *Annals of Eugenics*, 16, 199-214.

Fisher, R. A. (1915). "Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population," *Biometrika*, 10, 507-521.

Hotelling, H. (1951). "A generalized t-test and measure of multivariate dispersion," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Los Angeles and Berkeley, 23-42.

Kshirsagar, A. M. (1961 a). "The non-central multivariate Beta distribution," *Annals of Mathematical Statistics*, 32, 104-111.

Kshirsagar, A. M. (1961 b). "The goodness of fit of a single (non-isotropic) hypothetical principal component," *Biometrika*, 48, 397-407.

Lawley, D. N. (1938). "A generalization of Fisher's Z test," *Biometrika*, 30, 180-187.

Pillai, K. C. S. (1955). "Some new test criteria in multivariate analysis," *Annals of Mathematical Statistics*, 26, 117-121.

Ruben, H. (1966). "Some new results on the distribution of the sample correlation coefficient," *Journal of the Royal Statistical Society*, B, 28, 513-525.

Williams, E. J. (1955). "Significance tests for discriminant functions and linear functional relationships," *Biometrika*, 42, 360-381.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) SOUTHERN METHODIST UNIVERSITY	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP UNCLASSIFIED

3. REPORT TITLE

Correlation Between Two Vector Variables

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

A. M. Kshirsagar

6. REPORT DATE March 4, 1969	7a. TOTAL NO. OF PAGES 16	7b. NO. OF REFS 10
8a. CONTRACT OR GRANT NO. N00014-68-A-0515 b. PROJECT NO. NR 042-260 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 26 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research

13. ABSTRACT

H. Ruben (1966) has suggested a simple approximate normalization for the correlation coefficient in normal samples, by representing it as the ratio of a linear combination of a standard normal variable and a chi variable to an independent chi variable and then using Fisher's approximation to a chi variable. This result is extended in this paper to a matrix, which in a sense is the correlation coefficient between two vector variables $\underline{x}$ and $\underline{y}$. The result is then used to obtain large sample null and non-null (but in the linear case) distributions of the Hotelling-Lawley criterion and the Pillai criterion in multivariate analysis. Williams (1955) and Bartlett (1951) have derived some exact tests for the goodness of fit of a single hypothetical function to bring out adequately the entire relationship between two vectors $\underline{x}$ and $\underline{y}$, by factorizing Wilks' Λ suitably. These factors are known as "direction" and "collinearity" factors, as they refer to the direction and collinearity aspects of the null hypothesis. In this paper, the other two criteria viz. the Hotelling-Lawley and Pillai criteria are partitioned into direction and collinearity parts and large sample tests corresponding to them are derived for testing the goodness of fit of an assigned function.